①

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-305710
(43)Date of publication of application : 22.11.1996

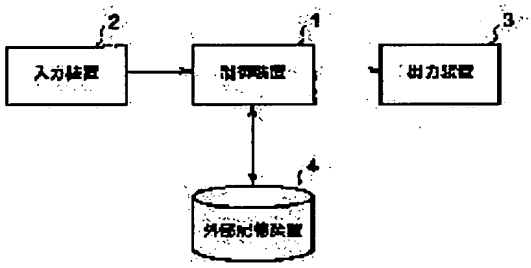
(51)Int.CI. G06F 17/30

(21)Application number : 07-106582
(22)Date of filing : 28.04.1995

(71)Applicant : TOSHIBA CORP
TOSHIBA COMPUT ENG CORP
(72)Inventor : OZAKI TOSHIHIRO
IWAI ISAMU

(54) METHOD FOR EXTRACTING KEY WORD OF DOCUMENT AND DOCUMENT RETRIEVING DEVICE

(57)Abstract:
PURPOSE: To automatically and effectively extract a keyword to be a key for a document.
CONSTITUTION: This document retrieving device is provided with a control device 1 consisting of a CPU for executing document keyword allocating processing and data processing and a memory, an input device 2 consisting of a keyboard or the like for inputting a document, a processing instruction, etc., an output device 3 consisting of a display or the like for dislaying a keyword allocated to a document and processing results, and an external storage device 4 such as an HDD for storing a data base or the like to be used for the allocation of a keyword and constituted so as to compare a certain document with another document stored in a document data base under the control of the control device 1 and extract a characteristic document keyword.



LEGAL STATUS

[Date of request for examination] 16.03.2000
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平8－305710

(43)公開日　平成８年(1996)11月22日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/30		9194－5L	Ｇ０６Ｆ　15/401	３１０Ａ	

審査請求　未請求　請求項の数10　ＯＬ　（全 14 頁）

(21)出願番号　特願平7－106582

(22)出願日　平成７年(1995)４月28日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(71)出願人　000221052
東芝コンピュータエンジニアリング株式会社
東京都青梅市新町1381番地１

(72)発明者　尾崎　敏宏
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内

(72)発明者　岩井　勇
東京都青梅市末広町２丁目９番地　株式会社東芝青梅工場内

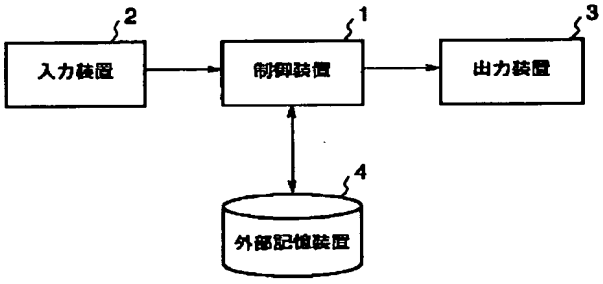
(74)代理人　弁理士　鈴江　武彦

(54)【発明の名称】　文書のキーワード抽出方法及び文書検索装置

(57)【要約】

【目的】本発明は、文書の鍵となるキーワードを自動的かつ効果的に抽出することを目的としたものである。

【構成】文書のキーワード付け処理やデータ処理を行なうＣＰＵやメモリからなる制御装置１と、文書や処理指示などを入力するキーボード等からなる入力装置１、文書に付与されたキーワードや処理結果を表示するディスプレイ等からなる出力装置３、文書のキーワード付けのためのデータベースなどを格納するＨＤＤ等の外部記憶装置４とを備え、制御装置１の制御の下に、文書データベースの他の文書と比較して特徴のある文書のキーワードを抽出することができる。



1

【特許請求の範囲】

【請求項１】　文書中に含まれる単語のうち、文書の要旨を知るための鍵となるキーワードを抽出する方法であって、

複数文書から文書中に含まれる単語を抽出し、

当該の単語を複数文書内での出現回数順に並び替えてランク付けし、

抽出した単語の種数で正規化した総文書単語ランク付けデータベースと、キーワード抽出対象文書に対して前記総文書単語ランク付けデータベースと同様の方法で作成した単語ランク付けデータベースを基にして、キーワード抽出対象文書から抽出した単語に関する総文書単語ランク付けデータベースと単語ランク付けデータベースでの各々のランクを求め、

その差分が許容された範囲以上に単語ランク付けデータベースでのランクが高い場合に、その単語がキーワード抽出対象文書のキーワードであると判断することを特徴とする文書のキーワード抽出方法。

【請求項２】　総文書単語ランク付けデータベースは、キーワード抽出対象文書の単語ランク付けデータベースの内容を追加して更新が可能なことを特徴とする請求項１記載の文書のキーワード抽出方法。

【請求項３】　複数の文書から抽出された単語毎に、複数文書の各文書に対してその単語の有無を記録した単語存在データベースを保持し、多くの文書に包含されている単語はキーワードとしないことを特徴とする請求項１又は２記載の文書のキーワード抽出方法。

【請求項４】　複数の文書から抽出した単語の各々に対して複数の文書内の出現回数をカウントして、その出現回数順に並び替えてランク付けを行ない抽出した単語の種数で正規化した総文書単語ランク付けデータベースを作成し、このデータベースを用いて文書のキーワードの抽出を行なう文書のキーワード付け装置であって、

キーワード抽出対象文書内の文字列を単語に分割する文書ー単語分割手段と、文書から分割した単語が文書中に何回出現するかをカウントする単語出現回数カウント手段と、

文書内の単語の出現回数で単語を並び替えてランク付けし抽出した単語の種数で正規化した単語ランク付け手段と、

文書から分割した単語について、総文書単語ランク付けデータベースと単語ランク付け手段とから夫々に得られたランクを比較して、ランクの差が指定された許容範囲を越える程度にキーワード抽出対象文書内に出現する頻度が高い場合に当該文書のキーワードと判断するキーワード判別手段と、

キーワードやその他の処理結果を出す出力手段とを具備したことを特徴とする文書のキーワード付け装置。

【請求項５】　総文書単語ランク付けデータベースは、キーワード抽出対象文書の単語ランク付けデータベース

2

の内容を追加して更新が可能なことを特徴とする請求項４記載の文書のキーワード付け装置。

【請求項６】　キーワード判別手段は、着目単語がキーワード抽出対象のキーワードか否かの判断をする際に、複数の文書から抽出された単語毎に、複数文書の各文書に対してその単語の有無を記録した単語存在データベースを保持し、多くの文書に包含されている単語はキーワードとしない単語存在データベースを用いて、どの文書にも現れるような単語をキーワードとしないことを特徴とする請求項４記載の文書のキーワード付け装置。

【請求項７】　文書を分野に分類する文書分類装置で、分類対象の分野毎にその分野を代表するキーワードを付与しておいた分野別キーワード情報を用いて文書を分野に分類する文書分類装置であって、

文書のキーワードを抽出するキーワード抽出手段と、

文書のキーワードと分野別キーワードを比較して一致した分野を当該文書の所属する分野であると判断する分野判断手段と、

文書を入力する入力手段と、結果を出力する出力手段とを具備したことを特徴とする文書分類装置。

【請求項８】　キーワード抽出手段は、請求項１記載の文書のキーワード抽出方法方法であることを特徴とする請求項７記載の文書分類装置。

【請求項９】　検索キーワードを含む文書をフルテキストサーチによって検索する文書検索装置であって、

検索キーの入力や処理指示を行なう入力手段と、

検索手段と、

検索結果を出力する出力手段と、

検索手段によって得られた検索キーワードを含む文書について検索キーワードの文書内での重要度によって並び替えを行って最終的な検索結果とする文書重要度判別手段とを具備することを特徴とする文書検索装置。

【請求項１０】　文書重要度判別手段は、検索の実行前に予め作成しておいた単語の種数で正規化した総文書単語ランク付けデータベース、及び単語ランク付けデータベースを基にして、２つのデータベースから得られる検索キーワードのランクの差を検索手段で求められた文書毎に数値化して重要度を判別することを特徴とする請求項９記載の文書検索装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、文書の要旨を知るための重要な単語をキーワードとして抽出する作業を支援し、文書を分類する作業を支援すると共に、文書データベース等に登録された文書をキーワードをもとに検索するようにした、文書のキーワード抽出方法と、キーワード付け装置、文書分類装置、及び文書検索装置に関する。

【０００２】

【従来の技術】従来、ある文書が何について述べている

3

かを分かりやすくするための手法には様々なものがある。一般的には、本文の要約や表題などで、また新聞等では、見出しを付与して内容を簡潔かつ効果的に伝える手法がとられている。また、文書の要旨を知るための鍵となる単語（キーワード）を取り出して、文書に付与する手法も存在する。

【０００３】文書にキーワードを付与する方法は、文書データベースに採用されている。文書の文書データベースへの登録時に、付与したキーワードも同時に登録する。この付与したキーワードは、文書データベースの検索時の検索キーワードとして用いられる。従来、このキーワードの付与作業は人手によって行なわれてきた。

【０００４】一方、機械的に文書中の単語を抽出するために、日本語解析や形態素解析を用いる方法も存在するが、この方法のみでは、抽出した単語が文書の要旨を知る鍵となるキーワードであるかの識別をつけることはできず、これを文書のキーワードとするには難点があった。

【０００５】また、文書が何の分野について述べられているかを判断し分類する作業については、効果的な分類を実施するためには文書の内容を理解しなければならず、これも人手によって分類が行なわれている場合が多い。

【０００６】また、文書の全文を検索対象としたフルテキストサーチでは、文書データベースに対して検索キーワードを含む文書を検索して得られた結果には、検索キーワードに関連が深い文書だけでなく、その検索キーワードが引用されているだけの相対的に重要度も低い文書も含まれていた。

【０００７】

【発明が解決しようとする課題】文書が何について書かれているものであるかを、文書の要旨を知るための鍵となるキーワードの形にする作業は人手によるものが多く、また、内容を理解するために、かなりの時間と専門性を必要とする。

【０００８】また、従来の機械的な文書からの単語の切り出しのみでは、文書のキーワードを抽出することができなかった。本発明は、上記事情を考慮して成されたものであり、文書の鍵となるキーワードを自動的かつ効果的に抽出することを目的としたものである。

【０００９】また、本発明は、これまで人手によって分類していた文書を前記基本発明である文書のキーワードの自動付与技術を用いて分類し、文書の分類作業の軽減を目的とする。

【００１０】また、フルテキストサーチでの検索支援があるが、文書の検索キーワードに対する有無を検索結果として出力するだけでなく、検索結果を検索キーの重要度が高いものについて並び替えて出力を行ない、ユーザーの要求する文書が探しやすくなる様な検索作業の効率化を目的とするものである。

4

【００１１】

【課題を解決するための手段】本発明は上記目的を達成するため、文書中に含まれる単語の内、文書の要旨を知るための鍵となるキーワードを抽出する方法であって、複数文書から文書中に含まれる単語を抽出し、当該の単語を複数文書内での出現回数順に並び替えてランク付けし、抽出した単語の種数で正規化した総文書単語ランク付けデータベースと、キーワード抽出対象文書に対して前記総文書単語ランク付けデータベースと同様な方法で作成した単語ランク付けデータベースを基にして、キーワード抽出対象文書から抽出した単語に関する総文書単語ランク付けデータベースと単語ランク付けデータベースでの各々のランクを求め、その差分が許容された範囲以上に単語ランク付けデータベースでのランクが高い場合に、その単語がキーワード抽出対象文書のキーワードであると判断することを特徴とする文書のキーワード抽出方法にある。

【００１２】また、本発明は上記目的を達成するため、複数の文書から抽出した単語の各々に対して複数の文書内の出現回数をカウントして、その出現回数順に並び替えてランク付けを行ない抽出した単語の種数で正規化した総文書単語ランク付けデータベースを作成し、このデータベースを用いて文書のキーワードの抽出を行なう文書のキーワード付け装置であって、キーワード抽出対象文書内の文字列を単語に分割する文書－単語分割手段と、文書から分割した単語が文書中に何回出現するかをカウントする単語出現回数カウント手段と、文書内の単語の出現回数で単語を並び替てランク付けし抽出した単語の種数で正規化した単語ランク付け手段と、文書から分割した単語について総文書単語ランク付けデータベースと単語ランク付け手段とから夫々に得られたランクを比較してランクの差が指定された許容範囲を越える程度にキーワード抽出対象文書内に出現する頻度が高い場合に当該文書のキーワードと判断するキーワード判別手段と、キーワードやその他の処理結果を出す出力手段とを具備したことを特徴とする文書のキーワード付け装置にある。

【００１３】また、本発明は上記目的を達成するため、文書を分野に分類する文書分類装置で、分類対象の分野毎にその分野を代表するキーワードを付与しておいた分野別キーワード情報を用いて文書を分野に分類する文書分類装置であって、文書のキーワードを抽出するキーワード抽出手段と、文書のキーワードと分野別キーワードを比較して一致した分野を当該文書の所属する分野であると判断する分野判断手段と、文書を入力する入力手段と、結果を出力する出力手段とを具備したことを特徴とする文書分類装置にある。

【００１４】更に、本発明は上記目的を達成するため、検索キーワードを含む文書をフルテキストサーチによって検索する文書検索装置であって、検索キーの入力や処

5

理指示を行なう入力手段と、検索手段と、検索結果を出力する出力手段と、検索手段によって得られた検索キーワードを含む文書について検索キーワードの文書内での重要度によって並び替えを行って最終的な検索結果とする文書重要度判別手段とを具備することを特徴とする文書検索装置にある。

【００１５】

【作用】上記構成に於いては、複数の文書から形態素解析等で単語切りを行って抽出した単語の複数文書に対する出現回数の順にランク付けを行ない出現した単語の種類の総数で正規化を行った基準となる総文書単語ランク付けベースと、キーワード抽出対象文書に対して前記の場合と同様に作成した単語ランク付けデータベースとを用いて、着目単語の２つのデータベースのランクの差を利用して対象文書のキーワードか否かの判定を行なうことにより、単純に文書を単語切りして抽出した単語がキーワードとはならず、また、２つのデータベースを相対的にに参照するために、文書内には頻繁に出現するが重要度が低いと判断される単語、例えば特許明細書における「手段」という単語は、対象文書のキーワードとして判定されることは無くなり、キーワード抽出対象文書の要旨を知る鍵となる的確なキーワードを抽出することが可能となる。

【００１６】また、上記構成に於いては、予め分野を代表するキーワードを分類したい分野毎に分野別キーワードとして付与しておくことで、分類対象文書から抽出したキーワードと一致する分野別キーワードを比較することが可能となり、一致した分野別キーワードを持つ分野に、分類対象文書を分類することが機械的に可能となり、人手を省略した効率的な分類が実現できる。

【００１７】更に、上記構成に於いては、予め作成しておいた総文書単語ランク付けデータベースと検索対象の各文書毎に作成した単語ランク付けデータベースを準備しておくことで、検索キーワードを含む文書の検索キーワードの重要度が数値的に判断可能となり、検索結果として検索キーワードの重要度に並べ替えを行った検索結果がユーザーに提示可能となり、検索の効率を向上することができる。

【００１８】

【実施例】本発明の概要は次の通りである。

（１）文書に含まれる単語が重要であるかの識別を行なう情報として、文書の鍵となる重要な単語は文書中に多く現れるという前提に着目した。また、文書中に多く現れるものの、さして重要ではない単語をキーワードとする事を防ぐため、文書データベース全体の情報を基にして、相対的に文書のキーワードを抽出する。詳しく述べるならば、複数の文書に現れる単語をその出現回数によってランク付けして単語種の総数で正規化した情報と、キーワードを抽出しようとする文書に現れる単語をその出現回数によってランク付けして単語種の総数で正規化

6

した情報の２つの情報からそれぞれにランクを取り出し、キーワードであるか否かを判断する際にそのランクの差分が誤差範囲と見なせない程、単語が数多く出現すると判断された場合に、当該文書のキーワードとする手法を採る。

【００１９】また、本発明は、文書を入力したり処理指示を与える入力手段と、文書を単語に分割する文書－単語分割手段と、文書から分割した単語が文書中に何回出現するかをカウントする単語出現回数カウント手段と、文書内の単語の出現回数を用いて計算した結果を基に単語をランク付けして単語種の総数で正規化した単語ランク付け手段と、複数の文書に対して文書中に含まれる単語毎に出現回数の総計を記録して出現回数を用いて計算した結果を基に単語をランク付けした総文書単語ランク付けデータベースと、分割した文書中の単語がキーワードとなる単語かを判断するキーワード判別手段と、キーワードやその他の処理結果を出す出力手段とを具備したことを特徴とする。

【００２０】また、前記総文書単語ランク付けデータベースは、キーワード抽出対象文書の単語ランク付けデータベースの内容を追加して更新が可能なことを特徴とする。また、前記キーワード判別手段は、検索実行前に予め作成しておいた、複数の文書から抽出された単語毎に各文書に対してその単語の有無を記録した単語存在データベースを保持し、多くの文書に含まれている単語はキーワードとしないことを特徴とする。

（２）文書を分野に分類する発明は、分類対象の分野毎にその分野を代表するキーワードを付与しておいた分野別キーワード情報と、文書のキーワードを抽出するキーワード抽出手段と、文書のキーワードと分野別キーワードを比較して一致した分野を当該文書の所属する分野であると判断する分野判断手段と、文書を入力する入力手段と、結果を出力する出力手段とを具備したことを特徴とする。

【００２１】また、前記のキーワード抽出手段は、

（１）の方法であることを特徴とする。

（３）文書検索を支援する発明は、検索キーの入力や処理指示を行なう入力手段と、検索手段と、検索結果を出力する出力手段と、検索手段によって得られた検索キーワードを含む文書について検索キーワードの文書内での重要度によって並び替えを行なって最終的な検索結果とする文書重要度判別手段とを具備したことを特徴とする。

【００２２】また、前記の文書重要度判別手段は、検索の実行前に予め作成しておいた（１）記載の総文書単語ランク付けデータベース、及び単語ランク付けデータベースを基にして、２つのデータベースから得られる検索キーワードのランクの差を検索手段で求められた文書毎に数値化して重要度を判別することを特徴とする。

【００２３】以下図面を参照して本発明の一実施例を説

明する。

（第１実施例）本実施例は、文書のキーワード付け装置に関わるものである。

【００２４】図１は、文書のキーワード付け装置に係わる概略構成を示すブロック図であり、１は文書のキーワード付け処理やデータ処理を行なうＣＰＵ、メモリ等からなる制御装置である。

【００２５】２は文書や処理指示などを入力するキーボード等からなる入力装置である。３は文書に付与されたキーワードや処理結果を表示するディスプレイ等からなる出力装置である。

【００２６】４は文書のキーワード付けのためのデータベースなどを格納するＨＤＤ等の外部記憶装置である。図２は、上記図１に示した制御装置１の詳細例を示したブロック図である。

【００２７】制御装置１は初期化部１１、入力部１２、出力部１３、制御部１４、文書－単語分割部１５、単語出現回数カウンタ部１６、単語ランク付け部１７、単語ランク積算部１８、キーワード判別部１９等の制御系と、文書格納バッファ２０、単語分割文書格納バッファ２１、単語登録バッファ２２、出現回数格納バッファ２３、キーワード格納バッファ２４等の記憶系と、メモリ上に確保した単語ランク付けデータベース２５、外部記憶装置に格納した総文書単語ランク付けデータベース２６等のデータベース系とから構成されている。

【００２８】初期化部１１は、記憶系の各バッファの初期化を行なう。入力部１２は入力装置２からデータベース作成モードか、キーワード抽出モードかのどちらかのモードの指定と文書の内容等の情報を入力する。出力部１３は入力部で指定されたモードがキーワード抽出モードだった時に、その入力された文書のキーワード等の情報を表示装置３に出力する。制御部１４は、制御系全体を制御して、データベース作成やキーワードの抽出などの各処理を総合的に制御する。

【００２９】入力部１２を通して入力されたキーワード抽出対象文書（以下文書と称する）は制御部１４を介して文書格納バッファ２０に格納される。文書－単語分割部１５では、日本語解析などの技術を用いて文書格納バッファ２０に格納された文書を単語に分割し、単語と単語の境がわかるようにして単語分割文書格納バッファ２１に格納し、さらに文書内で分割された単語を、重複することなく単語登録バッファ２２に格納する。

【００３０】単語出現回数カウンタ部１６は、単語登録バッファ２２に登録された単語毎に、該単語が単語分割文書格納バッファ２１における出現回数をカウントし、その結果を出現回数格納バッファ２３に格納する。単語ランク付け部１７はキーワード抽出モードの場合に動作し、出現回数格納バッファ２３の単語をその出現回数順に並び替え、出現回数を文書に出現した単語種の総数で正規化したものを単語ランク付けデータベース２５に格納する。

【００３１】単語ランク積算部１８は、データベース作成モードの場合に動作し、出現回数格納バッファ２３に出現した単語の内、未登録の単語を総文書単語ランク付けデータベース２６に格納し、登録済みの単語の場合は出現回数を加算して出現回数のデータを更新した後、該データベースに登録されている単語種の総数で再度正規化を行なって外部記憶装置４に格納する。

【００３２】キーワード判別部１９はキーワード抽出モードの場合に動作し、単語ランク付けデータベース２５に格納された単語の各々に対して該データベースと総文書単語ランク付けデータベース２６のランクを比較し、その差が閾値を越える程度に出現回数が多いと判断された場合に、該単語をキーワード格納バッファ２４に格納する。キーワード格納バッファ２４に格納されたキーワードは制御部１４を介して出力装置３に出力される。

【００３３】次に、本文書のキーワード付け装置の動作と処理の流れについて、図３に示すフローチャートを参照しながら説明する。まず、初期化部１１がステップ３０１にて起動し、各バッファを初期化する。次に、ステップ３０２にて、入力部１２によってモード（キーワード抽出モード、又はデータベース作成モード）の指示を受け、さらに入力部１２を介して入力された文書を、制御部１４が文書格納バッファ２０に格納する。図４は、文書格納バッファ２０に格納した文書の例を示した図である。

【００３４】文書格納バッファ２０に格納された文書は、ステップ３０３にて起動した文書－単語分割部１５によって単語単位に区切られ、区切り符号と共に単語分割文書格納バッファ２１に格納される。

【００３５】また、ステップ３０４では文書－単語分割部１５によって、抽出した単語を単語登録バッファ２２に格納する。図５は、単語分割文書格納バッファ２１の例であり、単語と単語の間にはスラッシュを区切り符号として用いてある。図６は、単語登録バッファ２２の例であり、文書から抽出した単語の例を示した図である。

【００３６】ステップ３０５では単語出現回数カウンタ部１６が起動して、単語分割文書格納バッファ２１と単語登録バッファ２２から、その抽出した単語毎に出現回数をカウントして、出現回数格納バッファ２３に格納する。図７は出現回数格納バッファ２３の例であり、単語の文字列とその出現回数を対応づけて格納してある。

【００３７】ステップ３０６ではステップ３０２で指定されたモードによって処理が異なる。キーワード抽出モードである場合はステップ３０７へ、データベース作成モードの場合はステップ３１２へ制御を移す。

【００３８】ステップ３０７では、単語ランク付け部１７が起動し、出現回数バッファ２３のデータを基にして、各単語を出現回数順に並び替え、さらに文書に出現した単語種の総数で正規化した値を付与して、単語ラン

ク付けデータベース２５として格納する。図８は、単語ランク付けデータベース２５の例であり、単語の文字列と出現回数と正規化値を対応づけて出現回数順に格納したものを示した図である。

【００３９】ステップ３０８～３１１ではキーワード判別部１９によって処理が行われる。ステップ３０８では、単語ランク付けデータベースの単語について、単語ランク付けデータベース２５と総文書単語ランク付けデータベースの該単語の正規化値を取り出し、この正規化値の差を求める。ステップ３０９では、ステップ３０８で求めた差が誤差の範囲を越える程に出現頻度が高いか否かを判断し、キーワードと判断された場合にはステップ３１０へ、そうでない場合はステップ３１１に制御を移す。

【００４０】ステップ３１０では、引き続きキーワード判別部１９によって処理が行われ、ステップ３０９でキーワードと判断された単語をキーワード格納バッファ２４に格納する。格納したキーワードは制御部１４を介して出力部１３に送られ、表示が行われる。

【００４１】ステップ３１１では、文書中の単語の全てについてキーワードの判定を行なったか否かを判断し、全てについて判定を行なっていない場合はステップ３０８に制御を移し、そうでない場合は処理を終了する。図９は、キーワード格納バッファ２４の例であり、抽出したキーワードを順に格納した図である。

【００４２】ステップ３１２～３１３はキーワード抽出モードの場合に、単語ランク付け積算部１８によって処理が行われる。ステップ３１２では出現回数格納バッファ２３に格納された単語の内、総文書単語ランク付けデータベース２６にない単語は、単語ランク付けデータベース２５の形で、登録済みの単語は単語ランク付けデータベース２５の出現回数を総文書単語ランク付けデータベース２６に加算する。ステップ３１３では、総文書単語ランク付けデータベース２６に出現する全ての単語を出現回数で並び替えを行ない、総単語種数で正規化を実施し、その正規化値とともに総文書単語ランク付けデータベース２６に格納して処理を終了する。図１０は、総文書単語ランク付けデータベース２６の例であり、単語ランク付けデータベース２５と同じフォーマットで格納したデータを示している。

（第２実施例）本実施例は文書の分類装置に関わるものである。

【００４３】文書分類装置に関わる概略構成を示すブロック図は第１実施例の図１と同様であり、１は文書の分類やデータ処理を行なうＣＰＵ、メモリ等からなる制御装置、２は文書や処理指示などを入力するキーボード等からなる入力装置、３は文書分類の結果を表示するディスプレイ等からなる出力装置、４は文書分類のための分野毎のキーワード等を格納するＨＤＤ等の外部記憶装置である。

【００４４】図１１は、図１に示した制御装置１の詳細例を示したブロック図である。制御装置１は、初期化部１１１、入力部１１２、出力部１１３、制御部１１４、キーワード抽出部１１５、文書分野判断部１１６等の制御系と、文書格納バッファ１１７、抽出キーワード格納バッファ１１８、分野別キーワード格納バッファ１１９等の記憶系と、外部記憶装置に格納した分野別付帯キーワード情報１２０等のデータベース系とから構成されている。

【００４５】初期化部１１１は、記憶系の各バッファの初期化を行ない、外部記憶装置４に格納された分野別付帯キーワード情報１２０を分野別キーワード格納バッファ１１９に格納する。

【００４６】入力部１１２は入力装置２から入力された文書を制御部１１４に渡す。出力部１１３は制御部１１４から渡された文書の分野等の処理結果を出力する。制御部１１４は文書分類装置の制御系全体を制御したり、入力部１１２からの文書データをバッファに格納する等の各処理を総合的に制御する。

【００４７】キーワード抽出部１１５は、文書格納バッファ１１７に格納された文書から文書のキーワードを抽出し、抽出キーワード格納バッファ１１８に格納する。文書分野別判断部１１６は、分野別キーワード格納バッファ１１９に格納された分野毎に予め登録されたキーワードと、抽出キーワード格納バッファ１１８に格納されたキーワードを比較し、一致したキーワードがある場合には該当分野を制御部１１４に渡す。

【００４８】次に、本文書の文書分類装置の動作と処理の流れについて、図１２に示すフローチャートを参照しながら説明する。ステップ１２０１では、初期化部１１１が動作を行ない、各バッファの初期化、及び外部記憶装置４の分野別付帯キーワード情報１２０を分野別キーワード格納バッファ１１９に展開する。図１３は、分野別付帯キーワード情報１２０と分野別キーワード格納バッファ１１９の格納例を示したもので、分野の名称とその分野に付帯するキーワードを列挙して格納した図である。

【００４９】次にステップ１２０２では、入力部が入力装置２から入力された文書を制御部１１４に渡し、制御部１１４は文書格納バッファ１１７に文書を格納する。文書格納バッファ１１７は第１実施例で示したものと同じで、この例は図４に示したものと同じである。

【００５０】ステップ１２０３ではキーワード抽出部１１５が動作を行ない、文書格納バッファ１１７に格納された文書のキーワードを抽出し、抽出キーワード格納バッファ１１８に格納する。抽出キーワード格納バッファ１１８は第１実施例で示したものと同じで、この例は図６で示したものと同じである。

【００５１】ステップ１２０４～１２０７までは文書分野別判断部１１６が動作を行なう。ステップ１２０４で

は、分野別キーワード格納バッファ１１９に格納した予め登録してあるキーワードと抽出キーワード格納バッファ１１８に格納したキーワードの調査を行ない、ステップ１２０５で一致しているものがあると判断した場合にはステップ１２０６に制御を移し、そうでない場合にはステップ１２０７に制御を移す。

【００５２】ステップ１２０６では一致したキーワードのある分野を制御部１１４に渡し、制御部１１４ではキーワードを出力部１１３に渡して出力装置３に対して出力を行なう。

【００５３】ステップ１２０７では、全ての分野について抽出したキーワードと一致するか否かのチェックを行ない、全ての分野についてチェックを終了していない場合にはステップ１２０４に制御を移し、そうでない場合には処理を終了する。

（第３実施例）本実施例は文書の検索装置に関わるものである。

【００５４】文書検索装置に係わる概略構成を示すブロック図は第１実施例の図１と同様であり、１は文書の検索や重要度を判断する制御を行なうＣＰＵ、メモリ等からなる制御装置、２は処理指示などを入力するキーボード等からなる入力装置、３は文書や検索の結果を表示するディスプレイ等からなる出力装置、４は検索対象文書や重要度を判別するデータ等を格納するＨＤＤ等の外部記憶装置である。

【００５５】図１１は、図１に示した制御装置１の詳細例を示したブロック図である。制御装置１は、初期化部１４０１、入力部１４０２、出力部１４０３、制御部１４０４、検索部１４０５、重要度判別部１４０６等の制御系と、検索中間結果格納バッファ１４０７、検索結果格納バッファ１４０８等の記憶系と、外部記憶装置に格納した検索対象文書１４０９、重要度判別データ等のデータベース系とから構成されている。

【００５６】初期化部１４０１は、記憶系の各バッファの初期化を行なう。入力部１４０２は入力装置２から入力された検索の指示や検索キーワード等を制御部１４０４に渡す。出力部１４０３は制御部１４０４から渡された検索結果等の処理結果を出力する。制御部１４０４は文書検索装置の制御系全体を制御したり、検索結果を出力装置に渡したり等の各処理を総合的に制御する。

【００５７】検索部１４０５は、外部記憶装置４に格納された検査対象文書１４０９から入力された検索キーワードを含む文書を検索し、該当文書に対応づけられた文書番号を検索中間結果格納バッファ１４０７に格納する。

【００５８】重要度判別部１４０６では、検索中間結果格納バッファ１４０７に格納された文書番号の文書について、外部記憶装置４に格納された重要度判別データ１４１０を用いて重要度のランク付けを行ない、重要度の順番に検索結果格納バッファ１４０８に文書番号を格納

する。

【００５９】次に、本文書の文書検索装置の動作と処理の流れについて、図１５に示すフローチャートを参照しながら説明する。ステップ１５０１では、初期化部１４０１が動作を行ない、各バッファの初期化を行なう。

【００６０】次にステップ１５０２では入力部１４０２が動作を行ない、入力装置２から入力された検索キーワードを制御部１４０４に渡す。ステップ１５０３では検索部１４０５が動作を行ない、制御部１４０４から渡された検索キーワードを含む文書を、検索対象文書１４０９から検索を行ない、該当文書の文書番号を検索中間結果格納バッファ１４０７に格納する。図１６は、検索中間結果格納バッファ１４０７の例であり、検索キーワードを含む文書の文書番号が格納されていることを示す図である。

【００６１】ステップ１５０４では重要度判別部１４０６が動作を行ない、検索中間結果格納バッファ１４０７に格納された文書番号の文書の重要度を重要度判別データ１４１０を用いて判断を行ない、文書の重要度の順に文書番号を検索結果格納バッファ１４０８に格納する。図１７は検索結果格納バッファ１４０８の例であり、検索キーワードを含む文書が重要度の順番に格納されていることを示す図である。

【００６２】ステップ１５０５では制御部１４０４が動作を行ない、検索結果格納バッファ１４０８に格納された文書番号を出力部１４０３に渡し、出力部１４０３は出力装置３へ文書番号の出力を行なう。

【００６３】ステップ１５０６では制御部１４０４が動作を行ない、入力部１４０２から入力された検索続行の指示の有無を基に制御を移行する。検索を続行する場合にはステップ１５０２に制御を移行し、そうでない場合には処理を終了する。

（第４実施例）本実施例は、文書のキーワード付け装置に関わるものである。

【００６４】本実施例のブロック図、制御装置の詳細ブロック図は第１実施例のものと同じであり、それぞれ、図１と図２が対応する。本実施例と第１実施例との差異は、図３を基にして述べるならば、ステップ３０６のキーワード抽出モードか否かの判断の後、ステップ３０７のキーワード抽出対象文書に現れる単語のランク付けのブロックの前に、ステップ３１２、ステップ３１３で述べた処理が挿入されるのみである。

（第５実施例）本実施例は、文書のキーワード付け装置に関わるものである。

【００６５】本実施例のブロック図は第１実施例のものと同じであり、図１が対応する。本実施例と第１実施例との差異は、図２を基にして述べるならば、総文書単語ランク付けデータベース２６を格納した外部記憶装置４に、複数の文書から抽出された単語毎に、複数文書の各文書に対してその単語の有無を記録した単語存在データ

ベースを保持することである。図１８は、単語存在データベースの格納例であり、行方向を単語文字列、列方向を文書番号として、マトリクス上に単語の文書に対する存在の有無を１、０で表した図である。

【００６６】この際、キーワード判別部１９は、実施例１で述べた、単語ランク付けデータベース２５に格納された単語の各々に対して該データベースと総文書単語ランク付けデータベース２６のランクを比較し、その差が閾値を越える程度に出現回数が多いと判断する処理に加えて、単語存在データベースを参照し多くの文書に含まれている単語はキーワードとしない処理を行ない、その結果をキーワード格納バッファ２４に格納する。

【００６７】また、処理の流れについて、第１実施例との差異を図３を基にして述べるならば、ステップ３０９のキーワードか否かの判断の後、ステップ３１０のキーワードの出力の前に、単語存在データベースを参照し多くの文書に含まれている単語はキーワードとしない処理が挿入される。

（第６実施例）本実施例は、文書分類装置に関わるものである。

【００６８】本実施例のブロック図は第２実施例のものと同じであり、図１が対応する。本実施例と第２実施例との差異は、図１１を基にして述べるならば、キーワード抽出部１１５が、第１実施例、第４実施例、第５実施例で述べた文書のキーワード抽出方法にて動作することである。

【００６９】また、図１２での第２実施例との差異は、ステップ１２０３での文書のキーワードを抽出する方法が、第１実施例、第４実施例、第５実施例で述べた文書のキーワード抽出方法にて処理が行われることである。

（第７実施例）本実施例は、文書検索装置に関わるものである。

【００７０】本実施例のブロック図は第３実施例のものと同じであり、図１が対応する。本実施例と第３実施例との差異は、図１４を基にして述べるならば、重要度判別部１４０６が、第１実施例、第４実施例、第５実施例で述べた文書のキーワード抽出方法にて動作することである。

【００７１】また、図１５での第３実施例との差異は、ステップ１５０４での重要度を判別する方法が、第１実施例、第４実施例、第５実施例で述べた文書のキーワード抽出方法にて処理が行われることである。

【００７２】

【発明の効果】以上詳記したように本発明によれば、文書データベースの他の文書と比較して特徴のある文書のキーワードを抽出することができる。また、本発明によれば、文書データベースの総文書単語ランク付けデータベースにキーワード抽出対象文書のデータを追加することで、登録時に総文書単語ランク付けデータベースの更新を行なうことができ、またデータ量が増えることにより、更に効果的なキーワードの抽出を行なうことができる。

【００７３】また、本発明によれば、単語存在データベースによってデータベース内の他の文書に多く現れる単語をキーワードとすることを防ぐことができる。また、本発明によれば、文書の分野分類作業を人手によって行なう必要がなくなり、効率の改善になる。

【００７４】また、本発明によれば、文書の分類作業を機械的に行なうため、文書の分類作業の性能が向上する。更に、本発明によれば、検索キーワードを含む文書の検索結果をユーザーに提示する際に、機械的に重要な文書をランクづけるために、ユーザーの検索の効率を向上させることができる。

【図面の簡単な説明】

【図１】本発明の実施例に係る装置の概略構成を示すブロック図。

【図２】本発明の第１実施例に於ける、文書のキーワード付け装置での図１の制御装置１の詳細な構成を示すブロック図。

【図３】本発明の第１実施例に於けるシステムの動作の概要を示すフローチャート。

【図４】本発明の第１実施例に於ける文書格納バッファ２０と第２実施例の文書格納バッファ１１７の格納例を示す図。

【図５】本発明の第１実施例に於ける単語分割文書格納バッファ２１の格納例を示す図。

【図６】本発明の第１実施例に於ける単語登録バッファ２２の格納例を示す図。

【図７】本発明の第１実施例に於ける出現回数格納バッファ２３の格納例を示す図。

【図８】本発明の第１実施例に於ける単語ランク付けデータベース２５の格納例を示す図。

【図９】本発明の第１実施例に於けるキーワード格納バッファ２４と第２実施例の抽出キーワード格納バッファ１１８の格納例を示す図。

【図１０】本発明の第１実施例に於ける総文書単語ランク付けデータベース２６の格納例を示す図。

【図１１】本発明の第２実施例に於ける文書分類装置での図１の制御装置１の詳細な構成を示すブロック図。

【図１２】本発明の第２実施例に於けるシステムの動作の概要を示すフローチャート。

【図１３】本発明の第２実施例に於ける分野別付帯キーワード情報１２０と分野別キーワード格納バッファ１１９の格納例を示す図。

【図１４】本発明の第３実施例に於ける文書検索装置での図１の制御装置１の詳細な構成を示すブロック図。

【図１５】本発明の第３実施例に於けるシステムの動作の概要を示すフローチャート。

【図１６】本発明の第３実施例に於ける検索中間結果格納バッファ１４０７の格納例を示す図。

15

【図１７】本発明の第３実施例に於ける検索結果格納バッファ１４０８の格納例を示す図。

【図１８】本発明の第５実施例に於ける単語存在データベースの格納例を示す図。

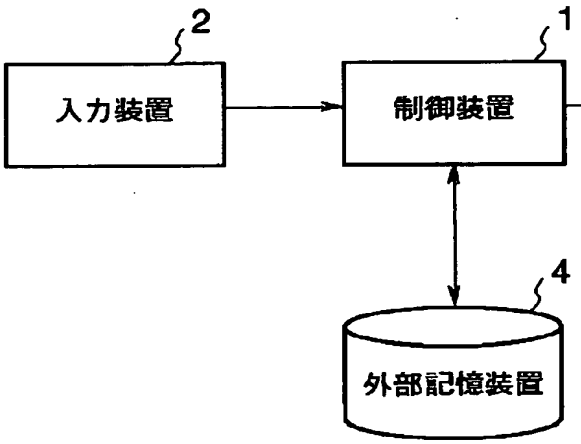
【符号の説明】

１…制御装置、２…入力装置、３…出力装置、４…外部記憶装置、１１…初期化部、１２……入力部、１３…出力部、１４…制御部、１５…文書ー単語分割部、１６…

16

単語出現回数カウント部、１７…単語ランク付け部、１８…単語ランク積算部、１９…キーワード判別部、１１１…初期化部、１１２…入力部、１１３…出力部、１１４…制御部、１１５…キーワード抽出部、１１６…文書分野判断部、１４０１…初期化部、１４０２…入力部、１４０３…出力部、１４０４…制御部、１４０５…検索部、１４０６…重要度判別部。

【図１】



【図６】

この 発明 は 、 半導体 に …

【図１６】

文書番号
7
24
39
146
255

【図４】

この発明は、半導体に関するもので AのB手段を具備する。…

【図５】

この/発明/は/、/半導体/に/関する/もの/で/ A/の/B/手段/を/具備/する/。/…

【図７】

出現回数	単語
18	この
43	発明
60	は
74	、
31	半導体
20	に

【図８】

出現回数	単語	正規化値
69	の	100.00
60	は	98.00
50	、	96.00
43	発明	94.00
42	手段	92.00
…	…	…
31	半導体	50.00
…	…	…

【図９】

半導体 製造

【図１０】

出現回数	単語	正規化値
125192	の	100.00
114837	に	98.04
103004	手段	88.52
…	…	…
5932	半導体	30.84

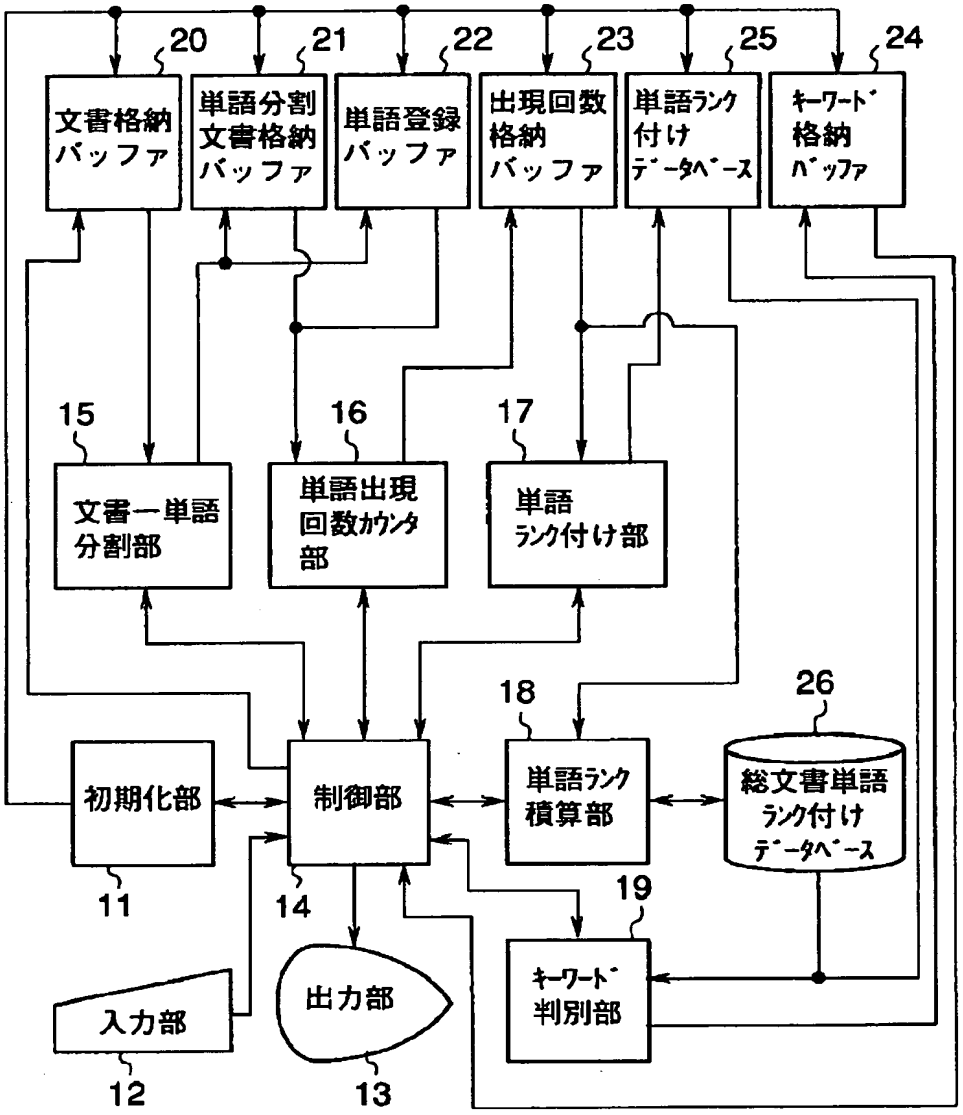
【図１３】

分野	キーワード
半導体	シリコン、ウェハ、IC、LSI、トランジスタ
メモリ	ROM、RAM、キャッシュ、PROM …
ネットワーク	LAN、WAN、WWW、インターネット…

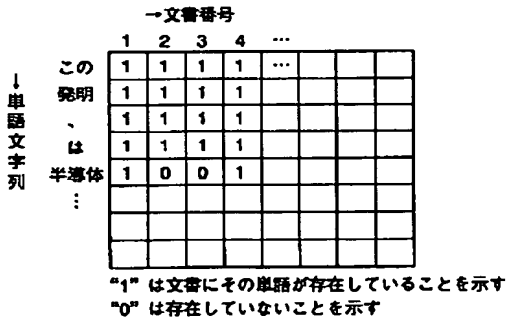
【図１７】

文書番号
146
24
7
255
39

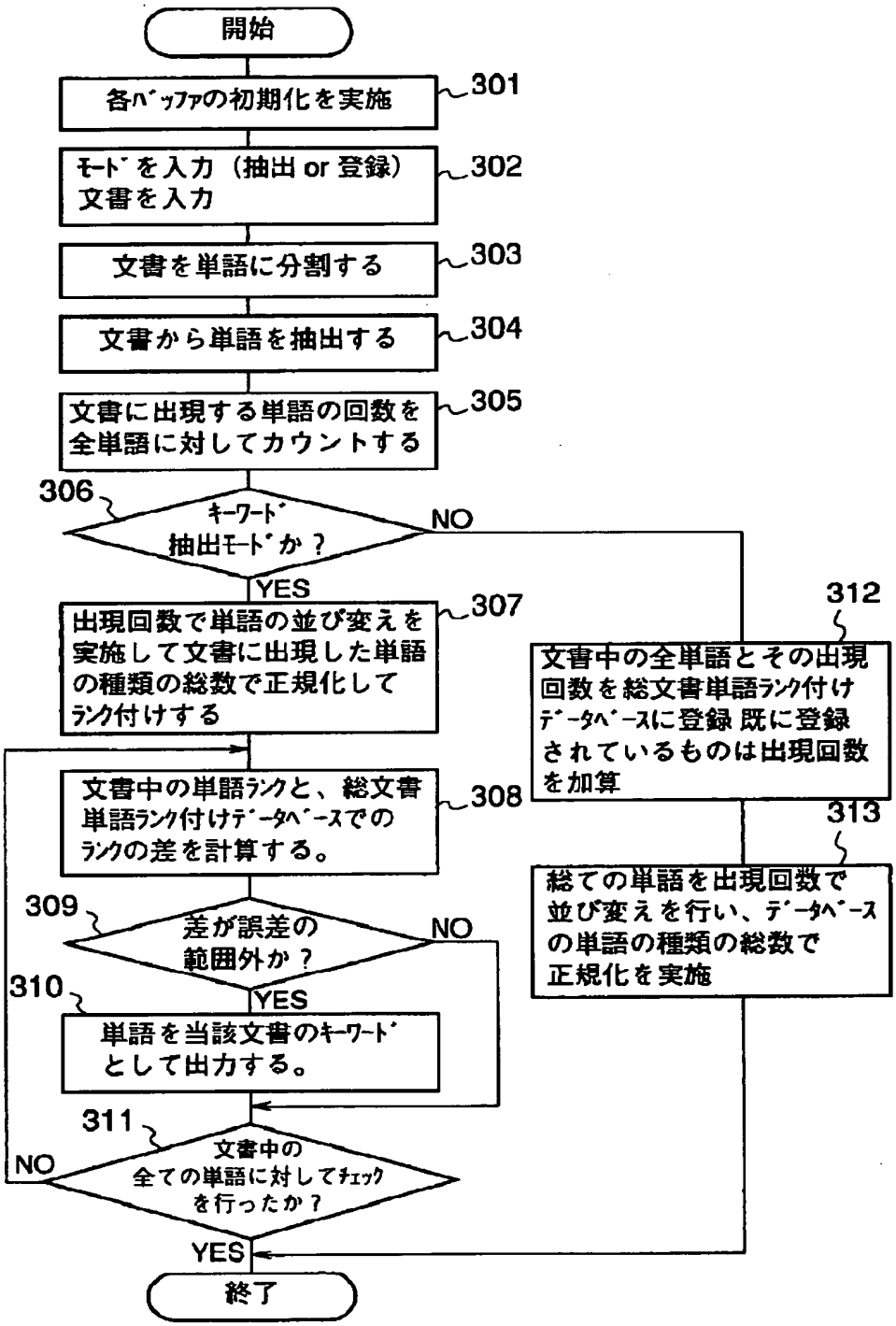
【図２】


20
21
22
23
25
24
文書格納バッファ
単語分割文書格納バッファ
単語登録バッファ
出現回数格納バッファ
単語ランク付けデータベース
キーワード格納バッファ
15
16
17
文書一単語分割部
単語出現回数カウンタ部
単語ランク付け部
18
26
初期化部
制御部
単語ランク積算部
総文書単語ランク付けデータベース
11
14
19
入力部
出力部
キーワード判別部
12
13

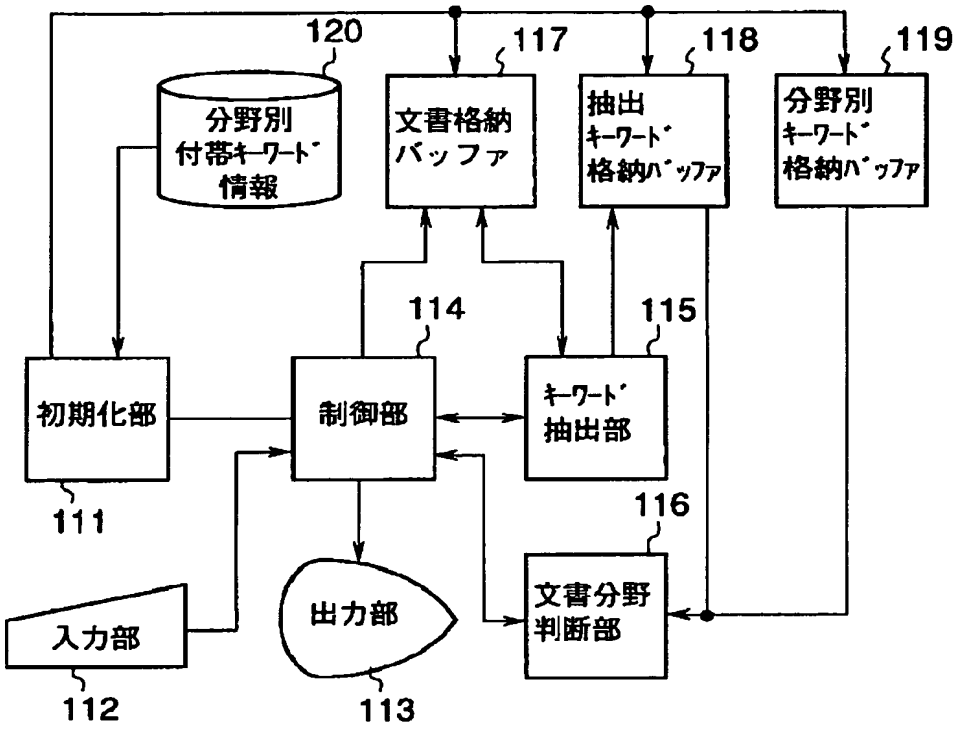
【図１８】


→文書番号
1 2 3 4 …
↓単語文字列
この
発明
、
は
半導体
…
"1" は文書にその単語が存在していることを示す
"0" は存在していないことを示す

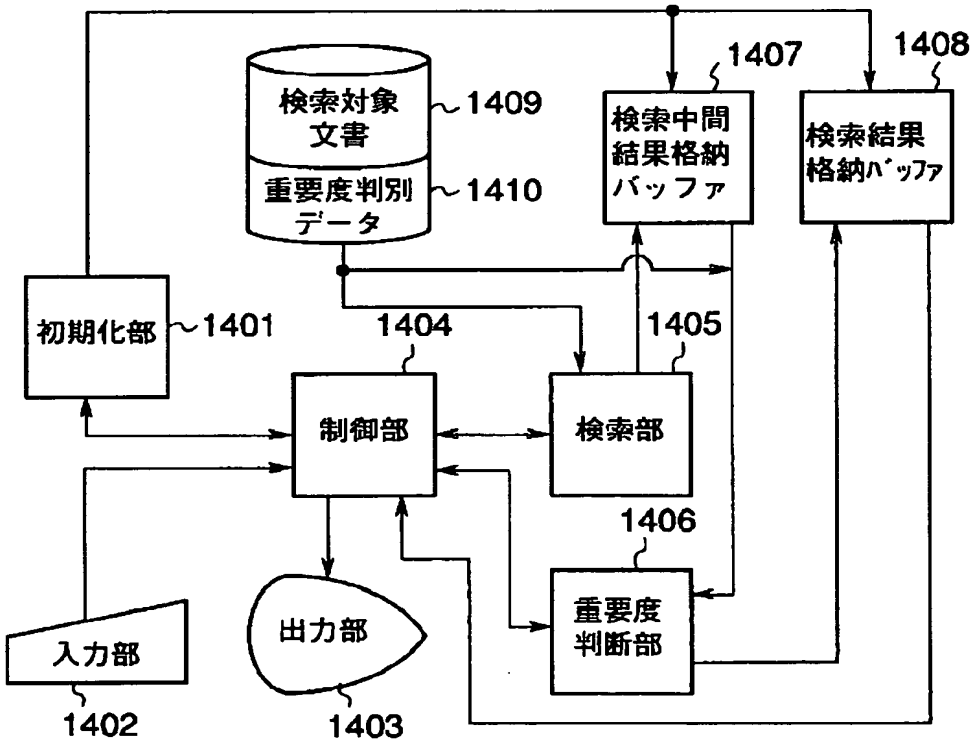
【図３】


開始
各ﾊﾞｯﾌｧの初期化を実施
301
ﾓｰﾄﾞを入力（抽出 or 登録）
文書を入力
302
文書を単語に分割する
303
文書から単語を抽出する
304
文書に出現する単語の回数を
全単語に対してカウントする
305
306
ｷｰﾜｰﾄﾞ
抽出ﾓｰﾄﾞか？
NO
YES
出現回数で単語の並び変えを
実施して文書に出現した単語
の種類の総数で正規化して
ﾗﾝｸ付けする
307
312
文書中の全単語とその出現
回数を総文書単語ﾗﾝｸ付け
ﾃﾞｰﾀﾍﾞｰｽに登録 既に登録
されているものは出現回数
を加算
313
総ての単語を出現回数で
並び変えを行い、ﾃﾞｰﾀﾍﾞｰｽ
の単語の種類の総数で
正規化を実施
文書中の単語ﾗﾝｸと、総文書
単語ﾗﾝｸ付けﾃﾞｰﾀﾍﾞｰｽでの
ﾗﾝｸの差を計算する。
308
309
差が誤差の
範囲外か？
NO
YES
310
単語を当該文書のｷｰﾜｰﾄﾞ
として出力する。
311
NO
文書中の
全ての単語に対してﾁｪｯｸ
を行ったか？
YES
終了

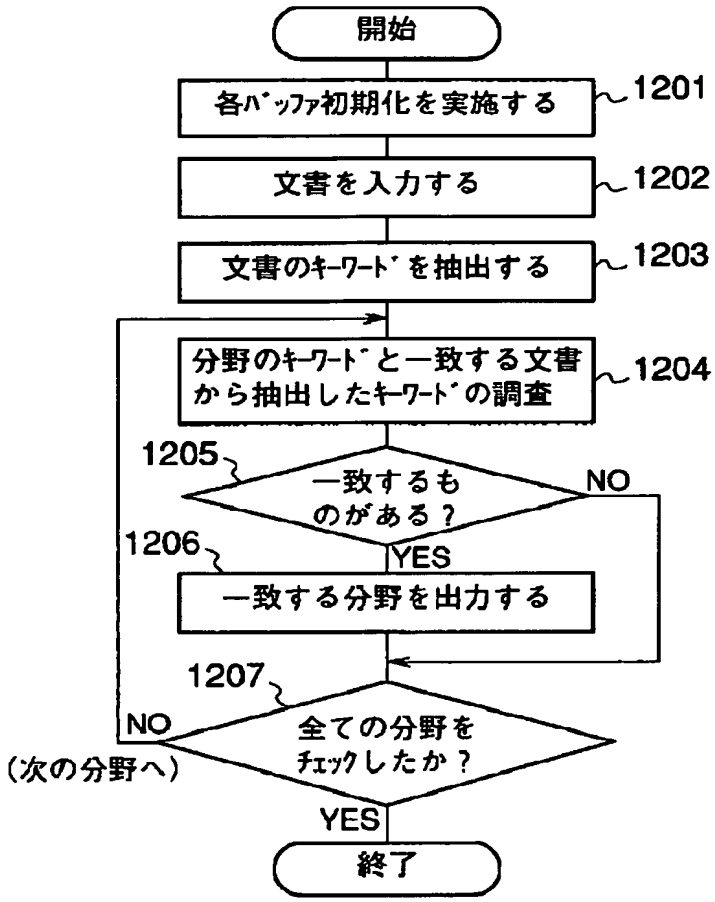
【図１１】


120
117
118
119
分野別
付帯ｷｰﾜｰﾄﾞ
情報
文書格納
ﾊﾞｯﾌｧ
抽出
ｷｰﾜｰﾄﾞ
格納ﾊﾞｯﾌｧ
分野別
ｷｰﾜｰﾄﾞ
格納ﾊﾞｯﾌｧ
114
115
初期化部
制御部
ｷｰﾜｰﾄﾞ
抽出部
111
116
入力部
出力部
文書分野
判断部
112
113

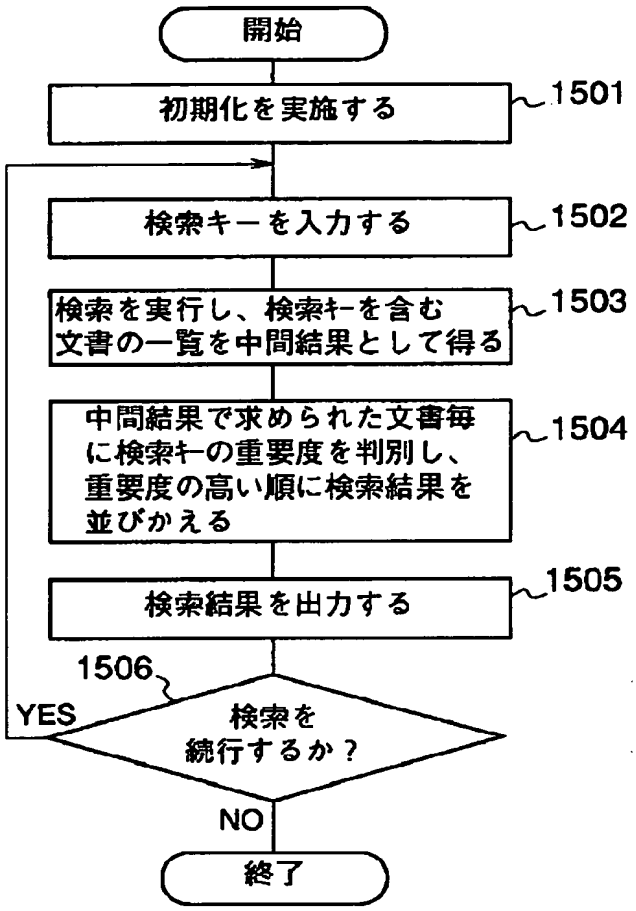
【図１４】


1407
1408
検索対象
文書
1409
検索中間
結果格納
ﾊﾞｯﾌｧ
検索結果
格納ﾊﾞｯﾌｧ
重要度判別
データ
1410
初期化部
1401
1404
1405
制御部
検索部
1406
重要度
判断部
入力部
出力部
1402
1403

【図１２】


開始
各ﾊﾞｯﾌｧ初期化を実施する
1201
文書を入力する
1202
文書のｷｰﾜｰﾄﾞを抽出する
1203
分野のｷｰﾜｰﾄﾞと一致する文書から抽出したｷｰﾜｰﾄﾞの調査
1204
1205
一致するものがある？
NO
YES
1206
一致する分野を出力する
1207
全ての分野をﾁｪｯｸしたか？
NO
(次の分野へ)
YES
終了

【図１５】


開始
初期化を実施する
1501
検索キーを入力する
1502
検索を実行し、検索キーを含む
文書の一覧を中間結果として得る
1503
中間結果で求められた文書毎
に検索キーの重要度を判別し、
重要度の高い順に検索結果を
並びかえる
1504
検索結果を出力する
1505
1506
YES
検索を
続行するか？
NO
終了